D562–D571 *Nucleic Acids Research, 2008, Vol. 36, Database issue* *Published online 18 October 2007*
doi:10.1093/nar/gkm758

CFGP: a web-based, comparative fungal genomics platform

Jongsun Park[1,2,3], Bongsoo Park[1,4], Kyongyong Jung[1,2], Suwang Jang[1], Kwangyul Yu[1], Jaeyoung Choi[1,2], Sunghyung Kong[1,2], Jaejin Park[1,2], Seryun Kim[1,2], Hyojeong Kim[3], Soonok Kim[3,5], Jihyun F. Kim[6], Jaime E. Blair[7], Kwangwon Lee[8], Seogchan Kang[4] and Yong-Hwan Lee[1,2,3,5,*]

[1]Fungal Bioinformatics Laboratory, [2]Department of Agricultural Biotechnology, [3]Center for Fungal Genetic Resource, Seoul National University, San 56-1, Sillim-9-dong, Seoul 151-921, Korea, [4]Department of Plant Pathology, The Pennsylvania State University, University Park, PA 16802, USA, [5]Center for Agricultural Biomaterials, Seoul National University, San 56-1, Sillim-9-dong, Seoul 151-921, Korea, [6]Systems Microbiology Research Center, Korea Research Institute of Bioscience and Biotechnology (KRIBB) 52 Oun-dong, Yuseong, Daejeon 305-806, Republic of Korea, [7]Department of Biology, Amherst College, Amherst, MA 01002 and [8]Department of Plant Pathology, Cornell University, Ithaca, NY 14853, USA

Received July 19, 2007; Revised September 9, 2007; Accepted September 11, 2007

ABSTRACT

Since the completion of the *Saccharomyces cerevisiae* genome sequencing project in 1996, the genomes of over 80 fungal species have been sequenced or are currently being sequenced. Resulting data provide opportunities for studying and comparing fungal biology and evolution at the genome level. To support such studies, the Comparative Fungal Genomics Platform (CFGP; http://cfgp.snu.ac.kr), a web-based multifunctional informatics workbench, was developed. The CFGP comprises three layers, including the basal layer, middleware and the user interface. The data warehouse in the basal layer contains standardized genome sequences of 65 fungal species. The middleware processes queries via six analysis tools, including BLAST, ClustalW, InterProScan, SignalP 3.0, PSORT II and a newly developed tool named BLASTMatrix. The BLASTMatrix permits the identification and visualization of genes homologous to a query across multiple species. The Data-driven User Interface (DUI) of the CFGP was built on a new concept of pre-collecting data and post-executing analysis instead of the 'fill-in-the-form-and-press-SUBMIT' user interfaces utilized by most bioinformatics sites. A tool termed Favorite, which supports the management of encapsulated sequence data and provides a personalized data repository to users, is another novel feature in the DUI.

INTRODUCTION

Fungi exert a far-reaching influence on the earth's biosphere (1). As recyclers of organic matter and as symbionts of most terrestrial plants, fungi are essential components of healthy ecosystems (2). For thousands of years, humans have exploited fungi for the production of many useful compounds and foods (3). In contrast to these benefits, fungi are also a major cause of plant diseases, significantly reducing crop yield (4). Fungi also represent a direct threat to human health as the most common cause of death in immunocompromised patients such as bone marrow transplant recipients and individuals suffering from advanced HIV infection due to systemic mycoses (5,6).

Studies on fungal biology have been greatly aided by rapidly accumulating genome sequence data (7). Since the completion of the genome sequencing of *Saccharomyces cerevisiae* (8), genomes of more than 80 fungal species have been completely sequenced or are currently being sequenced (7,9). As new high-throughput and low cost sequencing technologies (10) become widely available, the rate of fungal genome sequencing will continue to accelerate. Currently available fungal genome sequences cover species in four out of the seven fungal phyla, including Ascomycota, Basidiomycota, Chytridiomycota and Microsprodia (11,12) (Table 1). These genome sequences provide novel opportunities for elucidating the evolutionary and genetic basis of many different fungal lifestyle features, such as pathogenesis, symbiosis and the ability to grow on diverse substrates (9,13,14), via the use of various functional genomic and informatic tools.

*To whom correspondence should be addressed. Tel: +82-2-880-4674; Fax: +82-2-873-2317; Email: yonglee@snu.ac.kr

Table 1. List of genome sequences stored in the data warehouse of the CFGP

Species	Size (Mb)	No. of ORFs	Source[a]	Reference
Eubacteria (Domain)[b]				
Actinobacteria (Phylum)				
Bifidobacterium longum	2.3	1727	NCBI	(34)
Streptomyces coelicolor A3(2)	8.7	7769	CBS	(35)
Streptomyces avermitilis MA-4680	9.0	7575	CBS	
Proteobacteria (Phylum)				
Escherichia coli K12	4.6	4311	NCBI	(36)
Pseudomonas fluorescens Pf-5	7.1	6137	NCBI	(37)
Eukaryota (Domain)				
Cryptophyceae (Kingdom)[c]				
Guillardia theta	0.7	627	CBS	(38)
Euglenozoa (Kingdom)[c]				
Leishmania infantum	34.7	3241	SGTC	(39)
Fungi (Kingdom)[d]				
Ascomycota (Phylum)				
Pezizomycotina (Subphylum)				
Botrytis cinerea	42.7	16 448	BI	
Sclerotinia sclerotiorum	38.3	14 522	BI	
Aspergillus clavatus	27.9	9119	BI	
Aspergillus fischerianus	32.6	10 403	BI	
Apsergillus flavus	36.8	12 587	BI	
Aspergillus fumigatus	28.8	9926	TIGR	(40)
Aspergillus nidulans	30.1	10 701	BI	(17)
Aspergillus oryzae	37.1	12 062	DOGAN	(41)
Aspergillus terreus	29.3	10 406	BI	
Aspergillus niger	37.2	11 200	JGI	
Coccidioides immitis RS	28.9	10 457	BI	
Coccidioides immitis H538.4	55.6		BI	
Coccidioides immitis RMSCC 2394.1	28.9		BI	
Coccidioides posadasii Silveria	27.4		BI	
Coccidioides posadasii RMSCC 3488	28.1		BI	
Histoplasma capsulatum	33.0	9349	BI	
Uncinocarpus reesii	22.3	7798	BI	
Chaetomium globosum	34.9	11 124	BI	
Fusarium graminearum PH-1	36.6	13 321	BI	(42)
Fusarium graminearum GZ3639[e]	15.1		BI	(42)
Fusarium oxysporum	61.4	17 608	BI	
Fusarium verticillioides	41.9	14 155	BI	
Fusarium solani	51.3	15 707	JGI	
Magnaporthe oryzae	41.6	12 841	BI	(43)
Neurospora crassa	39.2	9822	BI	(44)
Podospora anserina	35.7	9872	IGM	
Trichoderma reesei	34.5	9997	JGI	
Alternaria brassicicola	32.0		WGSC	
Pyrenophora tritici-repentis	38.0		BI	
Mycosphaerella graminicola	41.9	11 395	JGI	
Mycosphaerella fijiensis	73.4	10 313	JGI	
Stagonospora nodorum	37.2	16 597	BI	
Saccharomycotina (Subphylum)				
Candida albicans SC5314	27.8	14 216	SGTC	(45)
Candida albicans WO-1	14.5	6157	BI	
Candida dubliniensis	14.5	6027	SI	
Candida glabrata	12.3	5174	CBS	(19)
Candida guilliermondii	10.6	5920	BI	
Candida lusitaniae	12.1	5941	BI	
Candida parapsilosis	13.1		SI	
Candida tropicalis	14.7	6258	BI	
Debaryomyces hansenii	12.2	6354	CBS	(19)
Eremothecium gossypii	8.7	4718	NCBI	(46)
Kluyveromyces lactis	10.7	5327	Genoscope	
Kluyveromyces waltii	10.6	5214	BI	(19)
Lodderomyces elongisporus	15.5	5796	BI	
Saccharomyces cerevisiae S288C	12.2	5898	SGD	(47)
Saccharomyces cerevisiae RM11-1a	11.7	5383	BI	
Saccharomyces cerevisiae YJM789	11.9	5471	SI	
Saccharomyces bayanus	11.5	9385	BI	(47)
Saccharomyces castellii	11.4	4677	VBI	(48)

(Continued)

Table 1. Continued.

Species	Size (Mb)	No. of ORFs	Source[a]	Reference
Saccharomyces kudriavzevii	11.2	3768	VBI	
Saccharomyces kluyveri	11.0	2968	WUGSC	(48)
Saccharomyces mikatae	11.5	9016	BI	(47)
Saccharomyces paradoxus	11.9	8939	BI	(47)
Pichia stipitis	15.4	5839	JGI	(49)
Yarrowia lipolytica	20.5	6524	CBS	(19)
Taphrinomycotina (Subphylum)				
Pneumocystis carinii[c]	6.3	4020	SI	
Schizosaccharomyces pombe	12.6	5005	GeneDB	(50)
Schizosaccharomyces japonicus	11.3	5172	BI	
Basidiomycota (Phylum)				
Agricomycotina (Subphylum)				
Postia placenta	90.9	17 173	JGI	
Phanerochaete chrysosporium	30.0	10 048	JGI	(51)
Coprinus cinereus	36.3	13 544	BI	
Laccaria bicolor	64.9	20 614	JGI	
Cryptococcus neoformans Serotype A	19.5	7302	BI	
Cryptococcus neoformans Serotype B	19.0	6870	NCBI	
Cryptococcus neoformans Serotype D B3501-A	19.3	6578	SGTC	(52)
Cryptococcus neoformans Serotype D JEC21	19.1	6475	SGTC	(52)
Pucciniomycotina (Subphylum)				
Sporobolomyces roseus	21.2	5536	JGI	
Puccinia graminis	88.7	20 567	BI	
Ustilaginomycotina (Subphylum)				
Ustilago maydis 521	19.7	6689	BI	(15)
Ustilago maydis FB1	19.7		BI	(15)
Chytridiomycota (Phylum)				
Batrachochytrium dendrobatidis	23.9	8818	BI	
Mucoromycotina (Subphylum *incertae sedis*)				
Rhizopus oryzae	45.3	17 467	BI	
Phycomyces blakesleeanus	55.9	14 792	JGI	
Microsporidia (Phylum)				
Encephalitozoon cuniculi	2.5	1996	Genoscope	(53)
Antonospora locustae	6.1	2606	JBPC	
Stramenopila (Kingdom)[c]				
Peronosporomycota (Phylum)				
Phytophthora infestans	228.5	22 658	BI	
Phytophthora sojae	86.0	19 276	JGI	(54)
Phytophthora ramorum	66.7	16 066	JGI	(54)
Hyaloperonospora parasitica	83.8		VBI	
Chloroplastida (Kingdom)[c]				
Charophyta (Phylum)				
Arabidopsis thaliana	119.2	28 581	TAIR	(55)
Oryza sativa var. japonica	370.8	37 555	IRGSP	(56)
Oryza sativa var. indica	426.3	49 710	BGI	(57)
Populus trichocarpa	485.5	58 036	JGI	(58)
Medicago truncatula	251.7	40 567	MTGSP	
Metazoa (Kingdom)				
Arthropoda (Phylum)				
Anopheles gambiae	287.8	15 802	Ensembl	(59)
Drosophila melanogaster	118.4	19 389	BDGP	(60)
Cnidaria (Phylum)				
Nematostella vectensis	356.6	27 273	JGI	(61)
Nematoda (Phylum)				
Caenohabditis elegans	100.3	21 124	NCBI	(62)
Urochordata (Phylum)				
Ciona intestinalis	173.5	19 744	Ensembl	(63)
Ciona savignyi	177.0	20 150	Ensembl	
Vertebrata (Phylum)				
Danio rerio	1636.5	14 966	Ensembl	
Tetraodon nigroviridis	402.2	28 005	Ensembl	
Xenopus tropicalis	1510.9	28 305	Ensembl	
Bos taurus	3144.2	32 991	Ensembl	
Canis familiaris	2519.8	30 308	Ensembl	
Gallus gallus	1105.2	24 166	Ensembl	
Pan troglodytes	4295.0	39 648	Ensembl	
Mus musculus	2724.2	36 471	Ensembl	(64)

(Continued)

Table 1. Continued.

Species	Size (Mb)	No. of ORFs	Source[a]	Reference
Rattus norvegicus	2718.9	32 543	Ensembl	
Homo sapiens	3418.7	33 869	Ensembl	(65)
Total	28 984.2	1 352 562		

[a]SGTC, Stanford Genome Technology Center; SI, Sanger Institute; CBS, Center For Biological Sequences; BI, Broad Institute; WGSC, Washington Univ. Genome Sequencing Center; JGI, DOE Joint Genomic Institute; DOGAN, Database Of the Genomes Analyzed at Nite; IGM, Instituté de Génétique et Microbiologie; TAIR, The Arabidopsis Information Resource; IRGSP, International Rice Genome Sequencing Project; BDGP, Berkeley Drosophila Genome Project; BGI, Beijing Genome Institute; VGI, Virginia Bioinformatics Institute; JBPC, Josephine Bay Paul Center for Comparative Molecular Biology and Evolution; MTGSP, Medicago Truncatula Genome Sequencing Project.
[b]Taxonomy based on (66).
[c]Taxonomy based on (67).
[d]Taxonomy based on (12).
[e]Incomplete coverage of genome information.

A better understanding of fungal biology will not only facilitate the judicious use of beneficial fungi, but also advance our efforts to control pathogenic species (15,16).

The abundance of sequenced species has facilitated in-depth comparative evolutionary genomic analyses across multiple fungal taxa (17–20). Because of the large amount of data involved, a cohesive, user-friendly informatics platform that links data and analysis tools is needed to efficiently support such analyses. Despite this need, the lack of data standardization has hampered the development of such platforms. The Genome Information Management System (GIMS) provided an integrated environment for archiving and visualization of genome sequences and data on transcriptome, protein–protein interaction, Gene Ontology (GO) and metabolic pathway (21). The 'eFungi', an improvement from the GIMS, stores genome sequences of 34 fungal and 2 Oomycete species (http://www.e-fungi.org.uk/). Although these systems systematically archive genomic data from multiple species, they do not support analysis of archived data with bioinformatic tools.

Heterogeneity of user interface (UI) and input/output data format in different bioinformatics tools has also complicated the integration of tools in a single platform to support multifaceted analyses of multiple genome sequences. Several systems provide multiple tools via a single platform. One example is the SNAP workbench, which supports sophisticated phylogenetic analyses through a menu-driven design (22). The iNquiry (BioTeam Inc., Wayland, MA, USA; http://web.bioteam.net/metadot/index.pl?id=2187) and European Molecular Biology Open Software Suite (EMBOSS) (23) are other examples of integrated, web-based platforms with multiple bioinformatic tools. The PLATCOM integrates a number of tools for comparative analysis of multiple genomes (24,25). These platforms, integrating data and tools, significantly shorten data analysis time by eliminating the need for visiting multiple, independent web sites to collect and analyse data. The ISYS platform utilizes middleware to link many different databases to data analysis tools using JAVA and allow these tools to communicate without any modification (26). Although these examples illustrate major improvements in supporting integrative analyses of genome sequence data via a single platform, the efficiency and expandability of such platforms require continuous enhancement, in order to adequately support utilization of rapidly increasing genome sequence data. Another area that requires improvement is the UI. Many currently available web-based bioinformatic platforms employ classical UI systems that simply display a list of functions or databases and provide a 'paste-sequence-and-press-submit' form (http://ausweb.scu.edu.au/aw02/papers/refereed/fitch/paper.html). Such UIs are easy to construct, but are not suitable for successively analysing sequence data with multiple tools.

To provide an effective means for analysing fungal genome sequence data through a suite of tools across multiple species, we developed the Comparative Fungal Genomics Platform (CFGP), which consists of a large-scale genomic data warehouse, bioinformatics tools useful for comparative genome analyses and a novel UI. The UI of the CFGP provides an easy access to sequence data stored in the data warehouse and seamlessly supports integrative data analyses using multiple tools. The data warehouse currently houses 101 genome databases in a standardized format for rapid data exchange. Bioinformatic tools incorporated into the CFGP were wrapped by a middleware program to efficiently manage tasks and facilitate data exchange between tools.

SYSTEM ARCHITECTURE AND DESIGN

The CFGP consists of three layers—the basal layer, middleware and the UI (Figure 1). The basal layer contains a data warehouse, which is managed using MySQL. Meta information for different types of biological data, including genome sequences, species and phenotype screening data, is placed as individual objects in this layer. The middleware connects the basal layer with the UI and supports the use of data analysis tools, including BLAST (27), ClustalW for multiple sequence alignment (28), InterProScan for predicting functional domains (29), SignalP 3.0 for predicting the presence of signal peptide (30), PSORT II for predicting subcellular localization (31) and a newly developed program named BLASTMatrix for identifying and summarizing the distribution pattern of homologous genes across the genome sequences stored in the CFGP. As a result of

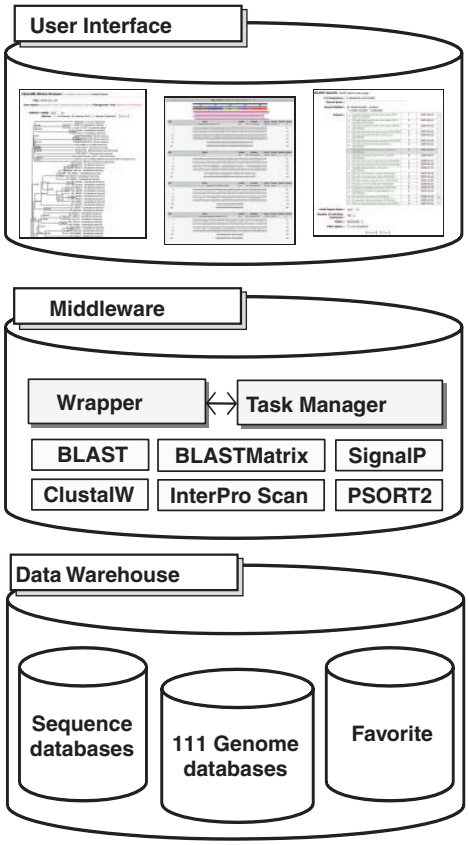


Figure 1. Overall system architecture and data flow in the CFGP. The basal layer contains a data warehouse, Favorite (a personal data repository and management tool), and external databases, such as InterPro and GO, stored in the CFGP. The wrapper in the middle layer relays requests from the UI to both the internal and external programs. The task manager at the right side of the wrapper manages tasks by assigning them to servers. At the upper layer, the DUI, a template engine developed with PHP, operates. A 'command' from the user goes to the middle layer. The basal layer passes the data to the middle layer as 'input'. At the middle layer, chosen programs generate results and pass them to the upper layer for 'representation' and to the basal layer for 'storage'.

the standardization of data exchange, the functionality of the CFGP can be easily expanded by adding any new tools that function in the UNIX environment. The UI of the CFGP developed with PHP (http://www.php.net) is based on a new concept, termed the Data-driven User Interface (DUI). By collecting sequences to be analysed first and executing analyses later, the DUI significantly reduces the time required for analysing the same sequence data via multiple tools.

The three layers of the CFGP can be manipulated and developed independently, which provides an optimal environment for maintenance and expansion of the CFGP. This was made possible by employing a standardized scheme in building each layer. In the basal layer, functions and schema of databases were standardized in both naming rules and basic structure of programming style, which enhances the efficiency of database development. In the middle layer, communications between the CFGP and external programs were standardized via PERL modules. This facilitates the future expansion of functionality, because new programs can be easily incorporated into the CFGP by constructing additional PERL modules. In the DUI, most of the interface components were standardized as a function so that a developer can easily make a new UI with selected components.

FEATURES OF THE CFGP

Data warehouse

Fungal genome sequence data in the public domain are stored in heterogeneous formats, posing a hurdle in integrating the data for comparative analysis. We retrieved these data and stored all Open Reading Frame (ORF) and contig (or chromosome) sequences of individual genomes in the data warehouse of the CFGP in a single format using MySQL. Subsequently, all sequence data were encapsulated as individual objects so that they can be easily analysed through multiple data analysis tools. The data warehouse currently houses the genome sequences of 65 fungal species, 4 Oomycete species and 27 non-fungal organisms (Table 1). The fungal genome databases cover 52 species belonging to the Ascomycota, eight species in the Basidomycota, two species each in the Mucoromycotina and the Microsprodia and one in the Chytridiomycota (12).

Data-driven user interface (DUI)

Most of the bioinformatics tools currently available through the web typically provide a box in the UI for pasting a query sequence. However, as the complexity of scientific inquires increases, often requiring multiple analyses with a single query, a single analysis with multiple sequences, or a combination of both, this type of UI becomes inefficient, and a new UI design is required (32). The only current solution for analysing a large number of sequences is a batch processing of data, which likely requires some level of programming knowledge by the user.

We developed the DUI to seamlessly support data management and integrative analyses using a suite of data analysis tools. It consists of two compartments: the Data Frame, supports browsing and collection of data, and the Manipulation Frame, which supports data management (Figure 2A). Four browsing tools under the 'SEQUENCE' menu include Contig Browser for browsing data in the data warehouse, SequenceSet Browser for browsing data in databases such as Uniprot, MyGene Browser for browsing data in the user's own computer and NR Browser for NR and NT sequences of NCBI. The Manipulation Frame provides a mechanism for storing and organizing data collected in a personalized space in the CFGP. The collection arrow transfers selected sequence data from the Data Frame to the Manipulation Frame, where they can be analysed by any bioinfomatic tools in the CFGP. This data management scheme significantly enhances the efficiency of data analysis, especially when large amounts of data are involved.

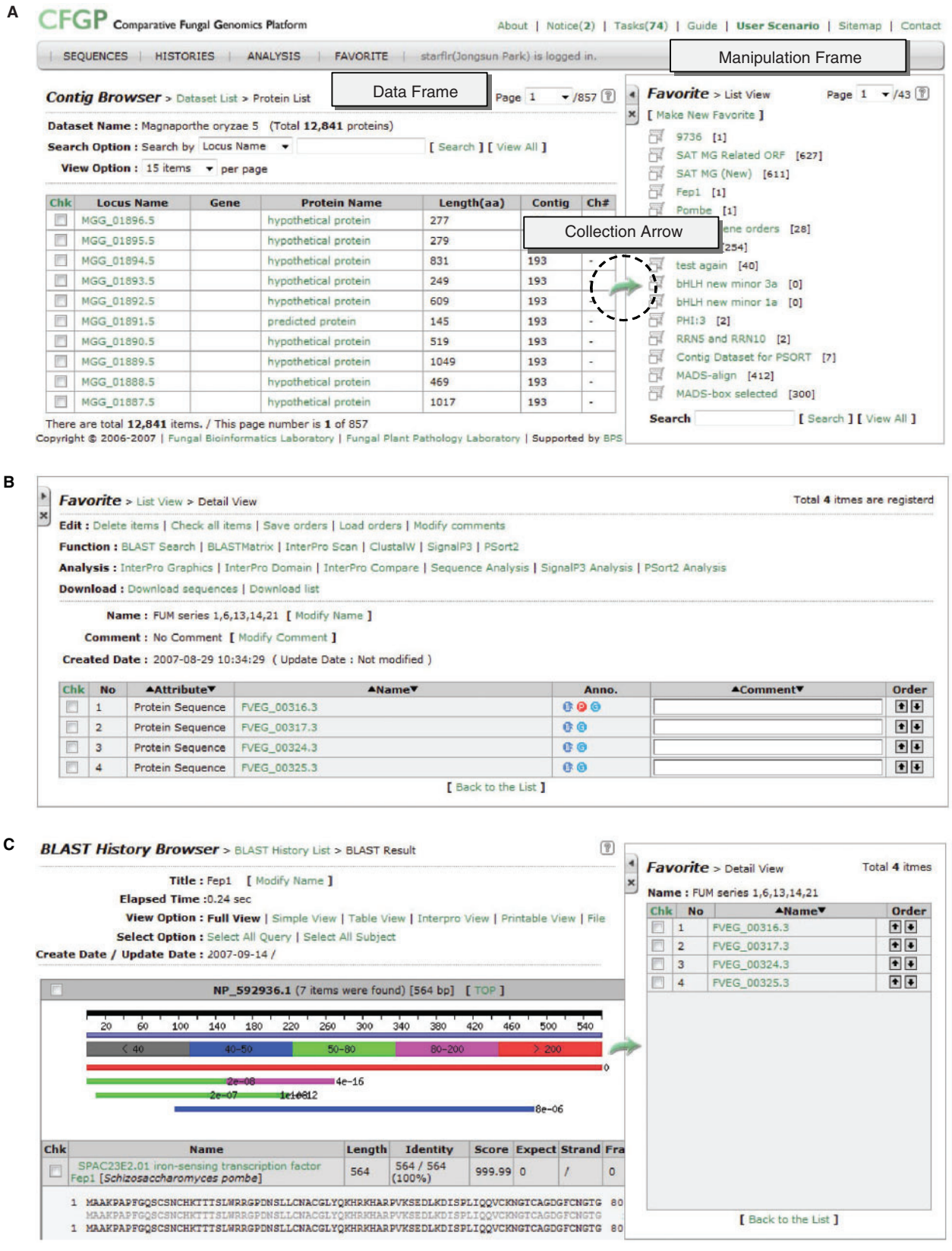
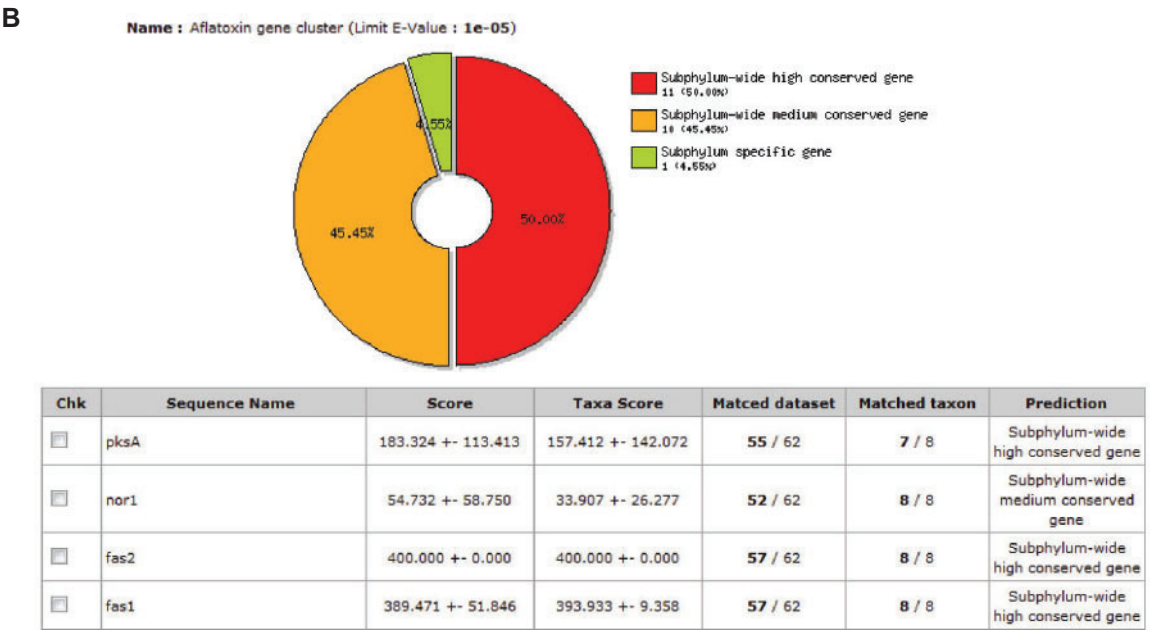


Figure 2. Structure of DUI. (**A**) A screenshot shows the process of data acquisition from Contig Browser. On the left side, 'Data Frame' displays the list of *Magnaporthe oryzae* proteins and 'Manipulation Frame' on the right side shows a list of Favorite. The 'Collection arrow' in the middle transfers chosen sequences from the Data Frame to the Manipulation Frame. (**B**) Collected sequences can be analysed by data analysis tools in Favorite. Users can choose sequences by clicking the checkbox in front of each sequence. (**C**) A BLAST search output is shown with Favorite in the Manipulation Frame. From the BLAST result, users can transfer sequences to Favorite via the use of the 'Collection Arrow'.



Chk	Sequence Name	Score	Taxa Score	Matced dataset	Matched taxon	Prediction
☐	pksA	183.324 +- 113.413	157.412 +- 142.072	**55** / 62	**7** / 8	Subphylum-wide high conserved gene
☐	nor1	54.732 +- 58.750	33.907 +- 26.277	**52** / 62	**8** / 8	Subphylum-wide medium conserved gene
☐	fas2	400.000 +- 0.000	400.000 +- 0.000	**57** / 62	**8** / 8	Subphylum-wide high conserved gene
☐	fas1	389.471 +- 51.846	393.933 +- 9.358	**57** / 62	**8** / 8	Subphylum-wide high conserved gene

Figure 3. Format of BLASTMatrix output. An example of BLASTMatirx output generated using the aflatoxin gene cluster in *Aspergillus nidulans* as queries. The results are presented in a matrix format (**A**) and a distribution based on e-value (**B**). Additionally, BLASTMatrix analyses the pattern of conservation in the BLASTMatrix dataset (such as novel gene, 'highly conserved gene' or 'taxon-specific gene') based on the distribution pattern of matched genes in all screened taxa.

Favorite as a bioinformatic workbench

A new UI tool named Favorite was developed to provide a personalized hub for storing and managing sequences retrieved from the data warehouse (Figure 2B). By storing only the primary keys of chosen sequences, not the sequences themselves, Favorite significantly reduces the space needed for storing data. Data stored in Favorite can be analysed with one tool or a series of tools by simply clicking the appropriate analyses in the option window (Figure 2C).

Five external programs, including BLAST, ClustalW, InterProScan, SignalP 3.0 and PSORT II, are available in Favorite. A BLAST search result can be presented in six different formats. One of them is 'interpro view', which displays the BLAST result annotated by InterPro to provide the functional prediction of the proteins in the

BLAST output. The ClustalW provides three different output formats: the multiple sequence alignment, distance matrix and the bootstrapped phylogenetic tree. The MSA viewer and Phyloviewer aid the user in manipulating the results of multiple sequence alignments and phylogenetic trees, respectively (http://phyloviewer.riceblast.snu.ac.kr; J. Park *et al.*, unpublished data). Results from InterProScan, SignalP and PSORT II are stored in the annotation database so that all results can be displayed in the annotation page of each query sequence. All analysis outputs provide an option of storing any sequences in the output into Favorite, offering an easy way to collect selected sequences for subsequent analyses.

To empower the personalized use of Favorite, user authentication is required. Besides supporting the management of individual users' data, Favorite can also be used to exchange data with other researchers. In addition, Favorite retains the user's original reference data, which overcomes any discrepancies between analyses conducted at different time points due to the frequent updating of external databases, such as the NR database in NCBI.

BLASTMatrix, a novel tool for searching and visualizing potential homologs across multiple species

With the availability of a large number of completely sequenced fungal genomes, it is possible to analyse the distribution of homologous genes across fungal taxa (7,9). Repeated BLAST searches against individual genome datasets are currently required for this task, which is iterative and cumbersome (33). To solve this problem, a new tool named the BLASTMatrix was developed and linked to the CFGP. With a query sequence, the BLASTMatrix generates a table containing the best hit in each of the species, which is then organized according to their taxonomical positions (Figure 3A), and also calculates the distribution pattern of homologous genes in different taxonomic groups (Figure 3B). The output can include InterPro or GO terms, helping the prediction of putative functions of hypothetical proteins. Further analyses can then determine the orthologous relationships between the query and its homologs in individual species.

FUTURE PROSPECTS

Genome sequences, along with associated functional genomics data, will continue to accumulate at an exponential rate. To efficiently utilize this inflow of data, standardization of data and efficient communication among data analysis tools are required. Enhancing the standard of communication between programs will also help future expansion by integrating more bioinformatics tools and will provide a development environment for open source projects. Additional genomic information, such as alternative splicing and expression data derived from EST, SAGE and microarray experiments, can be added to the CFGP.

ACKNOWLEDGEMENTS

This research was partially supported by grants from Crop Functional Genomics Center (CG1141) and Microbial Genomics and Applications Center (0462-20060021) of the 21st Century Frontier Research Program funded by the Ministry of Science and Technology and a grant from Biogreen21 Project (20050401034629) funded by Rural Development Administration to Y.H.L. A grant from the USDA-NRI Plant Biosecurity program (2005-35605-15393) to S.K. also supported this project. J.P. thanks to graduate fellowship provided by the Ministry of Education through the Brain Korea 21 Agricultural Biotechnology Project. Funding to pay the Open Access publication charges for this article was provided by the Brain Korea 21 Agricultural Biotechnology Project.

Conflict of interest statement. None declared.

REFERENCES

1. Hawksworth,D.L. (1991) The fungal dimension of biodiversity: magnitude, significance, and conservation. *Mycol. Res.*, **95**, 641–655.
2. Van der Heijden,M.G.A., Klironomos,J.N., Ursic,M., Moutoglis,P., Streitwolf-Engel,R., Boller,T., Wiemken,A. and Sanders,I.R. (1998) Mycorrhizal fungal diversity determines plant biodiversity, ecosystem variability and productivity. *Nature*, **396**, 69.
3. Demain,A.L. (2000) Microbial biotechnology. *Trends Biotechnol.*, **18**, 26–31.
4. Agrios,G.N. (2005) edn. *Plant Pathology, 5th edn*. Academic Press, San Diego.
5. Denning,D.W. (1998) Invasive aspergillosis. *Clin. Infect. Dis.*, **26**, 781–805.
6. Kwon-Chung,K.J., Varma,A. and Howard,D.H. (1990) Ecology of *Cryptococcus neoformans* and prevalence of its two varieties in AIDS and non-AIDS associated Cryptococcosis. In Vanden Bossche,H., Mackenzie,D.W.R., Cauwenbergh,G., Van Custem,J., Drouhet,E. and Dupont,B. (eds), *Mycoses in AIDS Patients*. Plenum Press, New York, pp. 103–113.
7. Galagan,J.E., Henn,M.R., Ma,L.J., Cuomo,C.A. and Birren,B. (2005) Genomics of the fungal kingdom: insights into eukaryotic biology. *Genome Res.*, **15**, 1620–1631.
8. Goffeau,A., Barrell,B.G., Bussey,H., Davis,R.W., Dujon,B., Feldmann,H., Galibert,F., Hoheisel,J.D., Jacq,C. *et al.* (1996) Life with 6000 genes. *Science*, **274, 546**, 563–547.
9. Park,J., Kim,H., Kim,S., Kong,S., Park,J., Kim,S., Han,H., Park,B., Jung,K. *et al.* (2006) A comparative genome-wide analysis of GATA transcription factors in fungi. *Genomics Inform.*, **4**, 147–160.
10. Metzker,M.L. (2006) Emerging technologies in DNA sequencing. *Genome Res.*, **15**, 1767–1776.
11. James,T.Y., Kauff,F., Schoch,C.L., Matheny,P.B., Hofstetter,V., Cox,C.J., Celio,G., Gueidan,C., Fraker,E. *et al.* (2006) Reconstructing the early evolution of fungi using a six-gene phylogeny. *Nature*, **443**, 818–822.
12. Hibbett,D.S., Binder,M., Bischoff,J.F., Blackwell,M., Cannon,P.F., Eriksson,O.E., Huhndorf,S., James,T., Kirk,P.M. *et al.* (2007) A higher-level phylogenetic classification of the Fungi. *Mycol. Res.*, **111**, 509–547.
13. Fitzpatrick,D.A., Logue,M.E., Stajich,J.E. and Butler,G. (2006) A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. *BMC Evol. Biol.*, **6**, 99.
14. Kroken,S., Glass,N.L., Taylor,J.W., Yoder,O.C. and Turgeon,B.G. (2003) Phylogenomic analysis of type I polyketide synthase genes in pathogenic and saprobic ascomycetes. *Proc. Natl Acad. Sci. USA*, **100**, 15670–15675.
15. Kamper,J., Kahmann,R., Bolker,M., Ma,L.J., Brefort,T., Saville,B.J., Banuett,F., Kronstad,J.W., Gold,S.E. *et al.* (2006)

Insights from the genome of the biotrophic fungal plant pathogen *Ustilago maydis*. *Nature*, **444**, 97–101.
16. Jeon,J., Park,S.Y., Chi,M.H., Choi,J., Park,J., Rho,H.S., Kim,S., Goh,J., Yoo,S. *et al.* (2007) Genome-wide functional analysis of pathogenicity genes in the rice blast fungus. *Nat. Genet*, **39**, 561–565.
17. Galagan,J.E., Calvo,S.E., Cuomo,C., Ma,L.J., Wortman,J.R., Batzoglou,S., Lee,S.I., Basturkmen,M., Spevak,C.C. *et al.* (2005) Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. oryzae*. *Nature*, **438**, 1105–1115.
18. Payne,G.A., Nierman,W.C., Wortman,J.R., Pritchard,B.L., Brown,D., Dean,R.A., Bhatnagar,D., Cleveland,T.E., Machida,M. *et al.* (2006) Whole genome comparison of *Aspergillus flavus* and *A. oryzae*. *Med. Mycol.*, **44**(Suppl.), 9–11.
19. Dujon,B., Sherman,D., Fischer,G., Durrens,P., Casaregola,S., Lafontaine,I., De Montigny,J., Marck,C., Neuveglise,C. *et al.* (2004) Genome evolution in yeasts. *Nature*, **430**, 35–44.
20. Kellis,M., Birren,B.W. and Lander,E.S. (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature*, **428**, 617–624.
21. Cornell,M., Paton,N.W., Hedeler,C., Kirby,P., Delneri,D., Hayes,A. and Oliver,S.G. (2003) GIMS: an integrated data storage and analysis environment for genomic and functional data. *Yeast*, **20**, 1291–1306.
22. Price,E.W. and Carbone,I. (2005) SNAP: workbench management tool for evolutionary population genetic analysis. *Bioinformatics*, **21**, 402–404.
23. Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, **16**, 276–277.
24. Choi,K., Ma,Y., Choi,J.H. and Kim,S. (2005) PLATCOM: a platform for computational comparative genomics. *Bioinformatics*, **21**, 2514–2516.
25. Lee,D., Choi,J.H., Dalkilic,M.M. and Kim,S. (2006) COMPAM: visualization of combining pairwise alignments for multiple genomes. *Bioinformatics*, **22**, 242–244.
26. Siepel,A., Farmer,A., Tolopko,A., Zhuang,M., Mendes,P., Beavis,W. and Sobral,B. (2001) ISYS: a decentralized, component-based approach to the integration of heterogeneous bioinformatics resources. *Bioinformatics*, **17**, 83–94.
27. McGinnis,S. and Madden,T.L. (2004) BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.*, **32**, W20–W25.
28. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
29. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bradley,P., Bork,P., Bucher,P. *et al.* (2005) InterPro, progress and status in 2005. *Nucleic Acids Res.*, **33**, D201–D205.
30. Bendtsen,J.D., Nielsen,H., von Heijne,G. and Brunak,S. (2004) Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.*, **340**, 783–795.
31. Nakai,K. and Horton,P. (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem. Sci.*, **24**, 34–36.
32. Wickware,P. (2000) Next-generation biologists must straddle computation and biology. *Nature*, **404**, 683–684.
33. Blair,J.E., Shah,P. and Hedges,S.B. (2005) Evolutionary sequence analysis of complete eukaryote genomes. *BMC Bioinformatics*, **6**, 53.
34. Schell,M.A., Karmirantzou,M., Snel,B., Vilanova,D., Berger,B., Pessi,G., Zwahlen,M.C., Desiere,F., Bork,P. *et al.* (2002) The genome sequence of *Bifidobacterium longum* reflects its adaptation to the human gastrointestinal tract. *Proc. Natl Acad. Sci. USA*, **99**, 14422–14427.
35. Bentley,S.D., Chater,K.F., Cerdeno-Tarraga,A.M., Challis,G.L., Thomson,N.R., James,K.D., Harris,D.E., Quail,M.A., Kieser,H. *et al.* (2002) Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature*, **417**, 141–147.
36. Blattner,F.R., Plunkett,G. 3rd, Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K. *et al.* (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1474.
37. Paulsen,I.T., Press,C.M., Ravel,J., Kobayashi,D.Y., Myers,G.S., Mavrodi,D.V., DeBoy,R.T., Seshadri,R., Ren,Q. *et al.* (2005) Complete genome sequence of the plant commensal *Pseudomonas fluorescens* Pf-5. *Nat. Biotechnol.*, **23**, 873–878.
38. Douglas,S.E. and Penny,S.L. (1999) The plastid genome of the cryptophyte alga, *Guillardia theta*: complete sequence and conserved synteny groups confirm its common ancestry with red algae. *J. Mol. Evol.*, **48**, 236–244.
39. Myler,P.J., Beverley,S.M., Cruz,A.K., Dobson,D.E., Ivens,A.C., McDonagh,P.D., Madhubala,R., Martinez-Calvillo,S., Ruiz,J.C. *et al.* (2001) The Leishmania genome project: new insights into gene organization and function. *Med. Microbiol. Immunol.*, **190**, 9–12.
40. Nierman,W.C., Pain,A., Anderson,M.J., Wortman,J.R., Kim,H.S., Arroyo,J., Berriman,M., Abe,K., Archer,D.B. *et al.* (2005) Genomic sequence of the pathogenic and allergenic filamentous fungus *Aspergillus fumigatus*. *Nature*, **438**, 1151–1156.
41. Machida,M., Asai,K., Sano,M., Tanaka,T., Kumagai,T., Terai,G., Kusumoto,K., Arima,T., Akita,O. *et al.* (2005) Genome sequencing and analysis of *Aspergillus oryzae*. *Nature*, **438**, 1157–1161.
42. Cuomo,C., Güldener,U., Xu,J., Trail,F., Turgeon,B., Di,P.A., Walton,J., Ma,L., Baker,S. *et al.* (2007) The *Fusarium graminearum* genome reveals a link between localized polymorphism and pathogen specialization. *Science*, **317**, 1400–1402.
43. Dean,R.A., Talbot,N.J., Ebbole,D.J., Farman,M.L., Mitchell,T.K., Orbach,M.J., Thon,M., Kulkarni,R., Xu,J.R. *et al.* (2005) The genome sequence of the rice blast fungus *Magnaporthe grisea*. *Nature*, **434**, 980–986.
44. Borkovich,K.A., Alex,L.A., Yarden,O., Freitag,M., Turner,G.E., Read,N.D., Seiler,S., Bell-Pedersen,D., Paietta,J. *et al.* (2004) Lessons from the genome sequence of *Neurospora crassa*: tracing the path from genomic blueprint to multicellular organism. *Microbiol. Mol. Biol. Rev.*, **68**, 1–108.
45. Jones,T., Federspiel,N.A., Chibana,H., Dungan,J., Kalman,S., Magee,B.B., Newport,G., Thorstenson,Y.R., Agabian,N. *et al.* (2004) The diploid genome sequence of *Candida albicans*. *Proc. Natl Acad. Sci. USA*, **101**, 7329–7334.
46. Dietrich,F.S., Voegeli,S., Brachat,S., Lerch,A., Gates,K., Steiner,S., Mohr,C., Pohlmann,R., Luedi,P. *et al.* (2004) The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. *Science*, **304**, 304–307.
47. Kellis,M., Patterson,N., Endrizzi,M., Birren,B. and Lander,E.S. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **423**, 241–254.
48. Cliften,P., Sudarsanam,P., Desikan,A., Fulton,L., Fulton,B., Majors,J., Waterston,R., Cohen,B.A. and Johnston,M. (2003) Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science*, **301**, 71–76.
49. Jeffries,T.W., Grigoriev,I.V., Grimwood,J., Laplaza,J.M., Aerts,A., Salamov,A., Schmutz,J., Lindquist,E., Dehal,P. *et al.* (2007) Genome sequence of the lignocellulose-bioconverting and xylose-fermenting yeast *Pichia stipitis*. *Nat. biotechnol.*, **25**, 319–326.
50. Wood,V., Gwilliam,R., Rajandream,M.A., Lyne,M., Lyne,R., Stewart,A., Sgouros,J., Peat,N., Hayles,J. *et al.* (2002) The genome sequence of *Schizosaccharomyces pombe*. *Nature*, **415**, 871–880.
51. Martinez,D., Larrondo,L.F., Putnam,N., Gelpke,M.D., Huang,K., Chapman,J., Helfenbein,K.G., Ramaiya,P., Detter,J.C. *et al.* (2004) Genome sequence of the lignocellulose degrading fungus *Phanerochaete chrysosporium* strain RP78. *Nat. Biotechnol.*, **22**, 695–700.
52. Loftus,B.J., Fung,E., Roncaglia,P., Rowley,D., Amedeo,P., Bruno,D., Vamathevan,J., Miranda,M., Anderson,I.J. *et al.* (2005) The genome of the basidiomycetous yeast and human pathogen *Cryptococcus neoformans*. *Science*, **307**, 1321–1324.
53. Katinka,M.D., Duprat,S., Cornillot,E., Metenier,G., Thomarat,F., Prensier,G., Barbe,V., Peyretaillade,E., Brottier,P. *et al.* (2001) Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature*, **414**, 450–453.
54. Tyler,B.M., Tripathy,S., Zhang,X., Dehal,P., Jiang,R.H., Aerts,A., Arredondo,F.D., Baxter,L., Bensasson,D. *et al.* (2006) Phytophthora genome sequences uncover evolutionary origins and mechanisms of pathogenesis. *Science*, **313**, 1261–1266.

55. AGI. (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.
56. IRGSP. (2005) The map-based sequence of the rice genome. *Nature*, **436**, 793–800.
57. Yu,J., Hu,S., Wang,J., Wong,G.K., Li,S., Liu,B., Deng,Y., Dai,L., Zhou,Y. *et al*. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). *Science*, **296**, 79–92.
58. Tuskan,G.A., Difazio,S., Jansson,S., Bohlmann,J., Grigoriev,I., Hellsten,U., Putnam,N., Ralph,S., Rombauts,S. *et al*. (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science*, **313**, 1596–1604.
59. Holt,R.A., Subramanian,G.M., Halpern,A., Sutton,G.G., Charlab,R., Nusskern,D.R., Wincker,P., Clark,A.G., Ribeiro,J.M. *et al*. (2002) The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science*, **298**, 129–149.
60. Kornberg,T.B. and Krasnow,M.A. (2000) The *Drosophila* genome sequence: implications for biology and medicine. *Science*, **287**, 2218–2220.
61. Darling,J.A., Reitzel,A.R., Burton,P.M., Mazza,M.E., Ryan,J.F., Sullivan,J.C. and Finnerty,J.R. (2005) Rising starlet: the starlet sea anemone, *Nematostella vectensis*. *Bioessays*, **27**, 211–221.
62. CSC. (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science*, **282**, 2012–2018.
63. Dehal,P., Satou,Y., Campbell,R.K., Chapman,J., Degnan,B., De Tomaso,A., Davidson,B., Di Gregorio,A., Gelpke,M. *et al*. (2002) The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. *Science*, **298**, 2157–2167.
64. Waterston,R.H., Lindblad-Toh,K., Birney,E., Rogers,J., Abril,J.F., Agarwal,P., Agarwala,R., Ainscough,R., Alexandersson,M. *et al*. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
65. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M. *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
66. Garrity,G.M. (2001) *Bergey's Manual of Systematic Bacteriology, 2nd edn*. New York, Springer.
67. Adl,S.M., Simpson,A.G., Farmer,M.A., Andersen,R.A., Anderson,O.R., Barta,J.R., Bowser,S.S., Brugerolle,G., Fensome,R.A. *et al*. (2005) The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. *J. Eukaryot. Microbiol*., **52**, 399–451.